Holistic assessment of call centre performance

Journal:	*IET Networks*
Manuscript ID	NET-2017-0018.R1
Manuscript Type:	Research Paper
Date Submitted by the Author:	26-Jun-2017
Complete List of Authors:	Smith, Edward; BT, ; Schormans, John
Keyword:	NETWORK MODELLING, PERFORMANCE EVALUATION, QUEUEING THEORY, SIMULATION

Holistic assessment of call centre performance

E.A.Smith[1*], J.A.Schormans[2]
[1]Orion Building, Adastral Park, Martlesham Heath, Ipswich , UK
[2]School of Electronic Engineering and Computer Sciences, Queen Mary University of London, London, UK
[*]Email: edward.a.smith@btinternet.com

Abstract: In modern call centres 60-70% of the operational costs come in the form of the human agents who take the calls. Ensuring that the call centre operates at lowest cost and maximum efficiency involves a trade-off of the cost of agents against lost revenue and increased customer dissatisfaction due to lost calls. Modelling the performance characteristics of a call centre in terms of the agent queue alone misses key performance influencers, specifically the interaction between channel availability at the media gateway and the time a call is queued. A blocking probability at the media gateway, as low as 0.45%, has a significant impact on the degree of queuing observed and therefore the cost and performance of the call centre.

Our analysis also shows how abandonment impacts queuing delay. However, the call centre manager has less control over this than the level of contention at the media gateway. Our commercial assessment provides an evaluation of the balance between abandonment and contention, and shows that the difference in cost between the best and worst strategy is £130K per annum, however this must be balanced against a possible additional £2.98 m exposure in lost calls if abandonment alone is used.

1. Introduction

Most Workforce Management (WFM) systems predict the performance characteristics of a group of agents using methods based on Erlang's queuing formula (Erlang C). However callers have a propensity to abandon [1] if they lose patience and this will reduce the degree of queuing and therefore the number of agents required to give acceptable service. Erlang C does not account for this and therefore tends to over-estimate the time taken to answer calls [1-2].

We argue that this approach of considering abandonment alone is flawed because it treats the agent queue as an isolated entity and not as part of an integrated system, where other capabilities may impact.

A simplified call centre architecture is shown in figure 1. We show that, when restrictions in the capacity of the media gateway, that forms the boundary between the telephony network and the call centre switching equipment, are examined, queuing is again less than predicted by Erlang C. We use both

analytical modelling and simulation to demonstrate that modelling the performance of a call centre in terms of the agent queue alone misses key performance influencers.

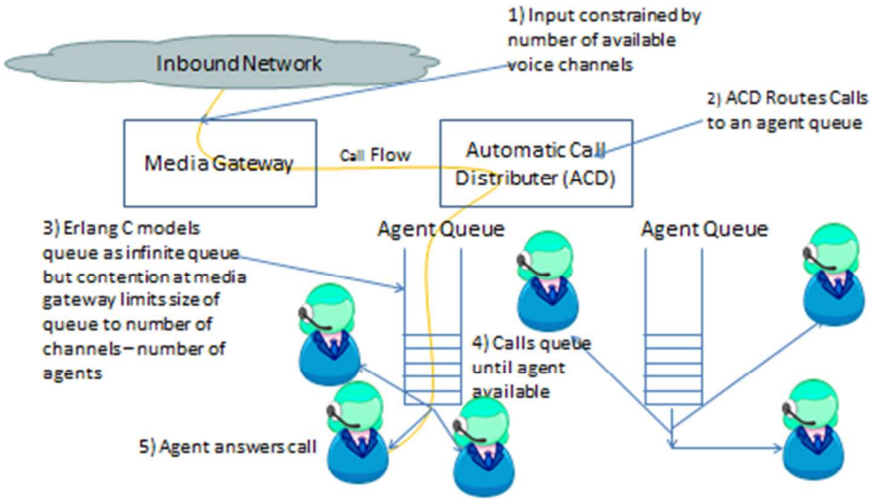

Figure 1: the basic architecture and call flow through a call centre.

We demonstrate that a grade of service (call blocking probability) at the media gateway as low as 0.45% has a significant impact. This constrains the size of the agent queue and reduces the probability of a call having to queue, significantly reducing the average call queuing time. This type of analysis could allow a call centre manager to more accurately judge the numbers of agents required to give the desired levels of service. We examine resulting financial implications later in this paper.

We contend that it is necessary to consider the input constraints (channels) as well as the output constraints (agents; often referred to in the literature as servers). Whilst agents are the largest costs, they are also the most flexible component of the system: varying the number of agents can be achieved in hours, but hardware changes such as adding channels takes weeks.

Our assessment will begin by considering recent work in the field, before describing our simulation model; we then give the results of our simulation modelled on a typical in-bound call centre. This will be done against two scenarios, one where there is a small amount of contention at the media gateway and one where there is none. We compare these results with analytical treatments. We then extend the analysis to

look at the behaviour of the call centre under circumstances where the load exceeds that normally experienced and moves into the realm of heavy traffic loading. We then consider the role of abandonment using both analytical models conforming to the Erlang A formula and through simulation.

We have also explored the impact of non-exponential service times. Finally we consider how varying system parameters impacts on the cost and effectiveness of the call centre.

2. Related Work

The contact centre is the primary interface between an organisation and its customers. Staff costs form 60-70% of the annual contact centre budget [1, 2] and many contact centres have a target agent utilisation of 90-92% [3]. Agents should therefore be deployed as effectively as possible. This is often achieved using management level planning tools, such as WFM, which use forecasted call arrival rates and assumed exponential distribution to identify the number of agents required to meet the demand within specified performance parameters and ultimately assess how resources are best utilised.

Two key measures of performance are: the degree to which calls are answered promptly and the level of call abandonment. Most WFM systems omit the impact of abandoned calls, which shortens the queue of callers waiting for an agent to become free and improves service for others [1, 2]. Strategies for managing performance optimise on the quality of the service to the customer, the efficient use of agents or a hybrid of the two [1, 4-7].

The queuing capability of the call centre may be assessed analytically using the Erlang C equation below:

$$P_c = \frac{N*A^N}{N!(N-A)} \frac{}{\sum_{i=0}^{N-1} \frac{A^i}{i!} + \frac{N*A^N}{N!(N-A)}} \qquad (1)$$

Here: N is the number of agents; A is the offered load in Erlangs and P_c is the queuing probability.

In Call Centre scenarios, equation 1 is an approximation, since it assumes infinite queue space. In the Call Centre case we know the queue can never get bigger than the difference between the number of channels

and the number of agents therefore this is inappropriate. If the maximum queue size is limited to Q, then the queuing probability is given by [8]:

$$P_c = \left(\frac{\frac{A^N}{N!}\left(1-\left(\frac{A}{N}\right)^{Q+1}\right)}{1-\frac{A}{N}} \frac{1}{\sum_{i=0}^{N-1}\frac{A^i}{i!}+\frac{\frac{A^N}{N!}\left(1-\left(\frac{A}{N}\right)^{Q+1}\right)}{1-\frac{A}{N}}} \right) \quad (2)$$

Taking T_s as the mean call service time, the average wait time, w is calculated as:

$$w = \frac{P_c\ T_s}{(N-A)} \quad (3)$$

This is the standard formula used in the Erlang C context. Under high load, the equations shown below [3] are often used instead of the conventional Erlang equations.

$$\beta = \sqrt{N}\,(1-\rho) \quad (4)$$

$$P_c = [1+\frac{\beta \boldsymbol{\Phi}(\beta)}{\emptyset(\beta)}]^{-1} \quad (5)$$

Where N is the number of agents, ρ is the utilisation, $\boldsymbol{\Phi}$ is the Normal probability integral and $\emptyset$ is the ordinate.

The role of abandonment has been explored extensively, including its impact when non-pre-emptive priority based queues are used [9] and the use of announcements to influence both balking and abandonment [10, 11]. Numerous studies have explored the impacts of various approaches to analytical modelling of abandonment [12-15].

Mandelbaum and Zeltyn [16] offer a method to calculate the wait probabilities for a multi-agent queue based on their treatment of Erlang A. Using their approach, the probability of waiting, when abandonment is taken into account is given by:

$$P_A = \frac{M\left(\frac{N}{\theta T_S}, \frac{\lambda}{\theta}\right) P_B}{1 + \left(M\left(\frac{N}{\theta T_S}, \frac{\lambda}{\theta}\right) - 1\right) P_B} \quad (6)$$

Where θ is the abandonment rate ($1/\theta$ is the average degree of patience exercised by the caller) and P_B is given by Erlang's loss formula. The term $M(x, y)$ is given below:

$$M(x, y) = 1 + \sum_{j=1}^{\infty} \frac{y^j}{\prod_{k=1}^{j} (x + k)} \quad (7)$$

Much of the current research focuses on inbound voice calls, but others address skills based routing and the potential to allow calls to overflow from one group to another under heavy load [17-22]. The channels available may extend beyond voice to multimedia techniques such as e-Mail and web chat, e.g. [23, 24].

Increasingly sophisticated queuing models have been evaluated including: the use of time varying inter-arrival rates [25, 26], heavy traffic approximations [27-31], the development of fluid models [32, 33], non-homogeneous Poisson arrivals and abandonment distributions [34-36] and the evaluation of service levels [37, 38]. Using data from Hydro Quebec and related simulation work [39], the impact of agent mix and experience on service time distribution has been explored. This analysis questions the use of exponentially and homogeneously distributed service times, given suggestions, based on a study of a small Israeli Bank, that service times may follow a log normal distribution [40].

Feinberg [41] looked at varying the number of lines serving a fixed number of agents and identified that the number of access circuits needs to exceed the number of agents by 10%. Increasing the number beyond this extends waiting times without increasing the fraction of served clients. He demonstrated that the fraction of served clients is independent of mean patience time and argued that lost clients and not lost calls characterise performance.

Massey and Wallace [42] analyse multi-server queues and demonstrate analytically that as the number of lines available is reduced, the level of blocking increases but the level of queuing diminishes. Weerasinghe and Mandelbaum [43] describe a trade-off between the impact of blocking and abandonment; both impact the cost of providing the contact centre. A mathematical analysis is produced, which balances the trade-off between the two mechanisms and determines the optimum compromise based on a cost function.

These observations suggest that queue attenuation, through either blocking or abandonment, can improve the caller experience. Control is invoked through a random process driven by: the caller's propensity to abandon in the case of abandonment or the call arrival pattern in the case of modest call blocking. This is paralleled by recent work from the data communications world, where probabilistic based discard through Active Queue Management techniques has been shown to improve the delay characteristics of multiplexed UDP streams [44].

3. Simulations in the normal operating range of 88 to 95% Agent Utilisation

We next describe our approach to modelling using the Riverbed Modeller tool set (https://www.riverbed.com/gb/products/steelcentral/steelcentral-riverbed), which has the advantage of a wide user base and the versatility to address problems beyond call centre performance. The model was built as a statistically based event driven model consisting of communicating finite state machines.

Calls are generated by a source module and directed into an emulated telephony network, which then routes the calls to a call distribution function. This in turn routes the calls, based on called number, to the appropriate process model representing the group of agents allocated to handling that type of call. This is a high level emulation of a modern soft PABX.

The model considers a medium sized call centre handling about 200 Erlangs of incoming traffic with exponentially distributed call holding times of on average 300 seconds. To support this workload we model the impact of varying the number of available agents in the range 210 to 227 corresponding to an agent utilisation between 88.1% and 95.2%. These numbers are close to those reported for a call centre in Charlotte [1] with a mean call inter-call gap of 1.3 seconds and mean call holding time of 307 seconds with 223 agents allocated during the busy period.

We examine two contention situations at the media gateway, one where 500 ports are configured and there is no call loss through channel blocking and the other where a modest level of contention of less than 0.45% is invoked by restricting the number of available channels to 240. When the level of blocking is increased beyond this, by either varying the number of agents available or by adjusting the call arrival rate, the results follow the pattern described in the heavy traffic portion of the paper.

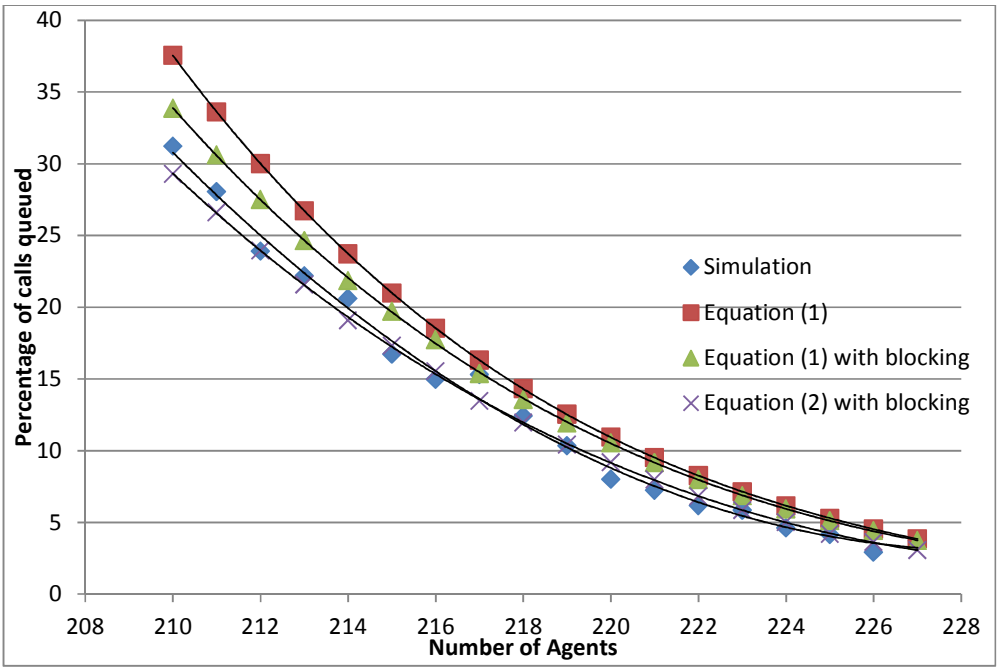

Figure 2 - the impact of agent numbers on call queuing when there is contention at the media gateway.

The analytical and DES results for the constrained queue are given in figure 2, which shows that the classic analytical result from equation 1 predicts more queuing than the simulation. Adjusting the value of A, in line with the blocking probability, gives the green curve in figure 2, which falls between the predictions of equations 1 and 2. Equation 2 gives a result that is very close to the simulation result. Thus the constraint at the media gateway reduces the degree of queuing in the system.

Figure 3 shows the time a call is queued for, showing simulation results alongside a number of analytical assessments based on equations 3 and 8 and the degree of queuing given by equations 1 and 2. The analytical result is initially very much larger than that delivered by the simulation, but the gap decreases as the number of agents increases. Modelling using the constrained form of the queuing probability yields a better fit with the simulation, but it still tends to over-estimate the queue time.

As described in the appendix, the system limits the number of ports into the system to *R* and therefore the maximum size of the queue becomes *R*-*N*, making the most likely mean queueing time limit $T_{max} = \frac{(\mathrm{R}-\mathrm{N})T_s}{N}$. This gives a more complex equation for the average wait time w:

$$w = P_c\left(\frac{T_s}{(N-A)} - \left(\frac{T_s}{(N-A)} + T_{max}\right)e^{-(N-A)T_{max}/T_s}\right) \qquad (8)$$

Figure 3 shows that the simulation results fall between those predicted by equations 3 and 8, when the degree of queuing is assessed using equation 2.

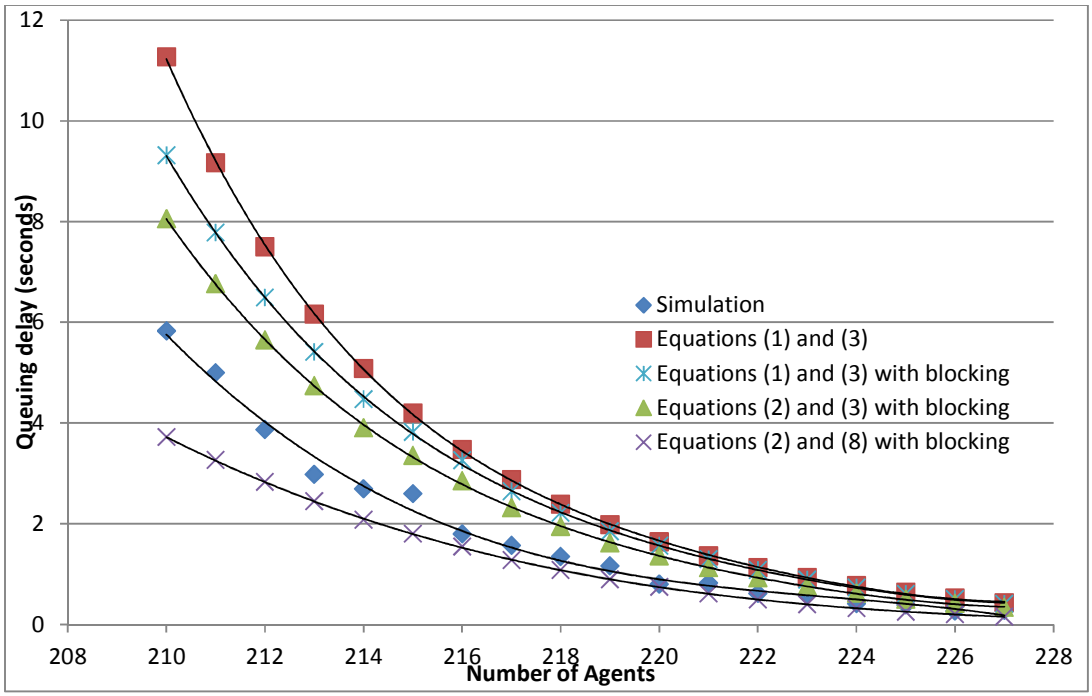

Figure 3 - the impact of agent numbers on mean call queuing delay in seconds with the queue constrained by contention at the media gateway.

These results suggest that contention at the media gateway constrains the agent queue size and hence reduces the probability of a call having to queue; more dramatically reducing the call queuing time.

When contention at the media gateway is removed, the analytical and simulation results are much closer together. If we assume that queuing space available is infinite, equation 2 simplifies to equation 1 and equation 8 simplifies to equation 3. While the maximum size for the queue is far from infinity, it is significantly larger than in the uncontended case, driving the term $(1 - \left(\frac{A}{N}\right)^{Q+1})$ very much closer to 1 and equation 2 towards the more approximate equation 1.

Examining the model results for both the constrained and unconstrained cases confirms that for a given population of agents the maximum queue size Q is significantly smaller when there is contention at the media gateway.

4. Simulation in the Heavy Traffic Range

We now consider what happens if the range of agents deployed is extended to the range of 155 to 240; that is a range that varies between where the load on the agents is greater than their capacity to deliver and where the load is reduced to 83% and the level of queuing drops to (virtually) zero.

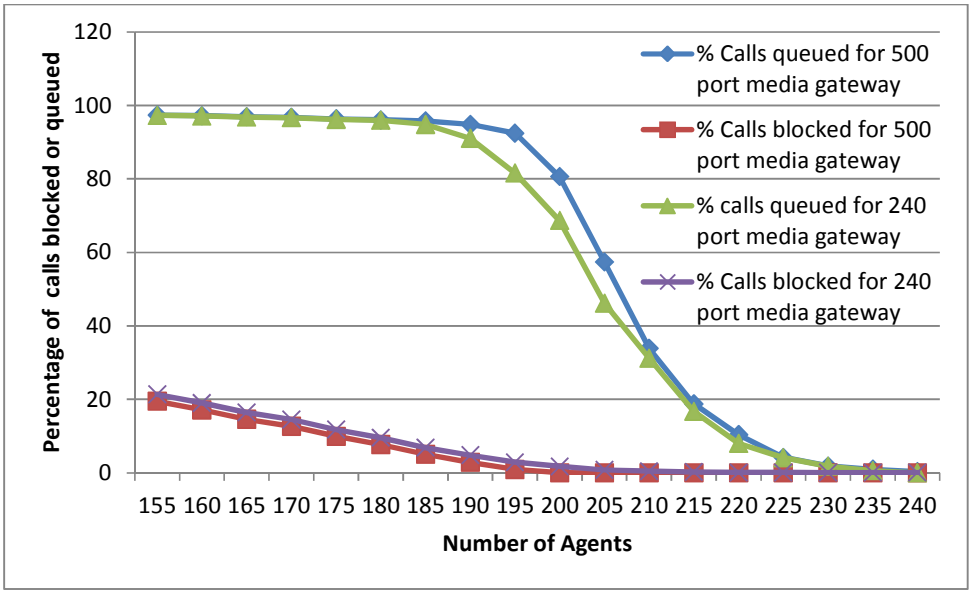

Figure 4 – Simulation results showing the variation in blocking at the media gateway and mean queuing in the agent queue.

The degree of blocking and percentage of calls queued are both high when the number of agents available is low and whilst the reduction in blocking as the number of agents is increased is gradual, the percentage of calls that are queued falls more dramatically as the number of agents increases beyond 185. This behaviour is echoed when queuing delay is examined.

A lower queuing delay is seen for the 240 port media gateway model, where more contention is expected, reflecting greater blocking and a lower limit on queue size due to the finite media gateway channel capacity.

At a low number of agents (less than 195) the queue shows high degrees of queuing. At these levels many authors resort to using heavy traffic approximations, such as those recommended by Halfin and Whitt [3]. We find that using such approaches do not simplify the computation process significantly and the results from incorporating both the heavy traffic and Erlang C evaluations are shown in figure 5.

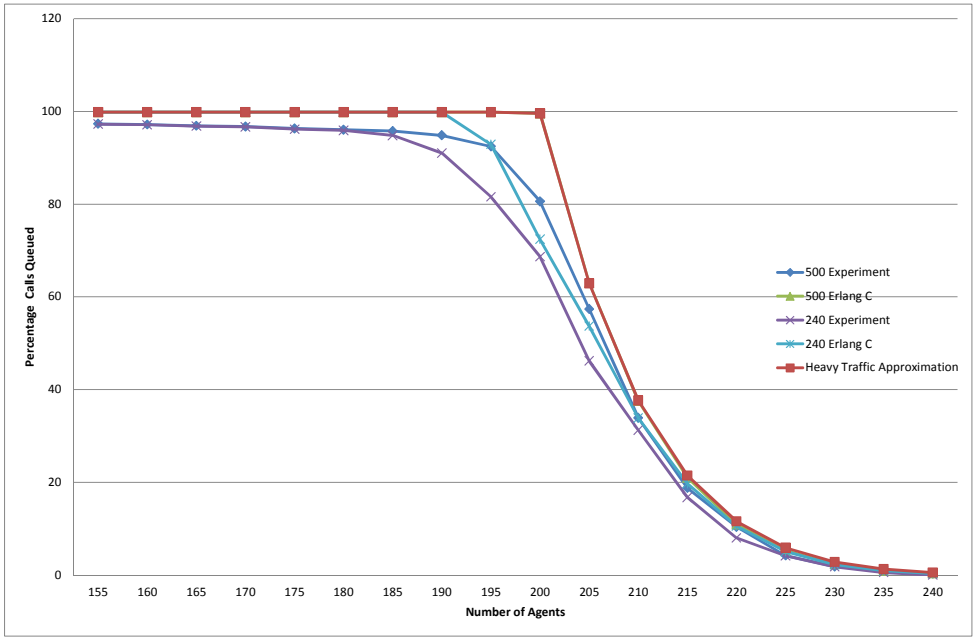

Figure 5 – Superimposing analytically derived heavy traffic behaviour on the queuing results returned by the simulation as the number of agents is varied from 155 to 240.

When the number of agents is less than 195, they are close to 100% loaded, the behaviour of the simulation fits that predicted by equation 5 and the Erlang C formula. As agents become busier the queue fills and call holding times increase, as does agent occupancy and hence the degree of call blocking. Blocking occurs once the number of agents plus number of calls queued are greater or equal to the number of available channels limiting the maximum usable queuing space. From this point new arrivals can only

be accepted at the same rate that calls are serviced, only when the number of agents is greater than 200 does the service rate exceed the arrivals rate.

When the number of agents is around 195, the results change rapidly as agent occupancy drops below 0.9999. The degree of queuing drops rapidly as the number of agents is increased beyond this level. There is good agreement between both of the analytical models and the simulation.

5. Abandonment

To simplify our analysis we have so far neglected abandonment, however given its importance we now consider its impact, which closely mirrors that of contention at the media gateway. Abandonment attenuates waiting times by reducing the average size of the queue, with those callers who have been held longest being the most likely to abandon.

The results obtained using equations 6 and 7 were used as the basis for developing an Excel spreadsheet, which enabled the easy production of graphical output. Looking explicitly in the range 210-227 agents, we examine performance as the patience time ($1/\theta$) is varied between 30 seconds and 5 minutes.

Increasing the patience time reduces the tendency to abandon calls. For a 30 second patience level, less than 2% of calls are abandoned, which is within the range recommended by Koole et al [1]. The impact of abandonment on queueing appears to be more marked than the effects due to contention at the media gateway. In the abandonment scenario no limit is placed on the size of the queue, but the outliers in the queue are pruned by the abandonment process. Whilst changing the patience time has a substantial impact on queuing times, its impact on the degree of abandonment is far smaller; a relatively aggressive patience time of 30 seconds gives an abandonment rate of approximately 2% for the 210 agent case.

We incorporated abandonment into our Riverbed model, through a minor modification to the queuing algorithm and discovered that there was very little difference between the simulations and the calculations derived from equations 6 and 7 until the patience time has extended to 300 seconds. The results are outlined in table 1.

Whilst the gap between the 240 ports (contended) simulation and the 500 port simulation seems to be extending, the ranges continue to overlap at the 95% level. Extending the analysis to a greater level than

300 seconds would make the level of patience the same as the service time configured and 50 times greater than the maximum wait delay experienced.

Patience time	Call Queuing Time			% Calls queued			% Calls abandoned		
	Analytical	Contended Simulation	Uncontended Simulation	Analytical	Contended Simulation	Uncontended Simulation	Analytical	Contended Simulation	Uncontended Simulation
30 S	0.59	0.56±0.07	0.56±0.07	12.98	11.7	11.7	1.96	1.86	1.86
60 S	1.023	1.02±0.12	1.02±0.12	16.26	15.24	15.24	1.705	1.63	1.63
120 S	1.696	1.63±0.25	1.60±0.21	19.9	19.93	19.82	1.41	1.46	1.46
180 S	2.225	2.09±0.32	2.08±0.27	21.12	20.68	21.02	1.24	1.15	1.17
240 S	2.66	2.20±0.32	2.58±0.42	23.68	20.78	21.31	1.11	0.97	1.02
300 S	3.04	2.76±0.40	3.27±0.48	24.88	22.84	22.11	1.02	0.88	0.86
Infinite	11.3	5.82±0.76	10.27±2.3	37.56	31.22	33.89	0	0	0

Table 1: Call queuing time, degree of call queuing and degree of abandonment as a function of patience time for a system served by 210 agents

We conclude that although abandonment and restriction of the queue size by contention at the media gateway both limit the degree of queuing, even at fairly high patience levels, the effect of the former quickly dwarfs that of the latter. In our judgement both serve to limit the number of calls being queued and therefore the call queuing time.

Extending the analytical treatment by increasing the range of agents from 210-227 to 155-240 changes the observed results markedly. At the 30 second patience level, with 155 agents, the degree of abandonment is high as expected (worst case 22%), as the protracted wait times mean that the more calls cross the patience threshold. The level of queuing follows a smooth curve as the number of agents is increased and becomes flatter at the edges and steeper in the middle portion as the user's patience increases approaching the behaviour predicted from the analytical Erlang C treatment and seen in the Riverbed simulation.

If the level of patience is low then abandonments will be at a sufficiently high level to reduce the degree of queuing and the use of the heavy traffic algorithm is no longer necessary.

6. Non-exponential Service Times

The literature suggests that hyper exponential service distributions should also be considered [45]. Their impact on call delays are shown in figure 6:

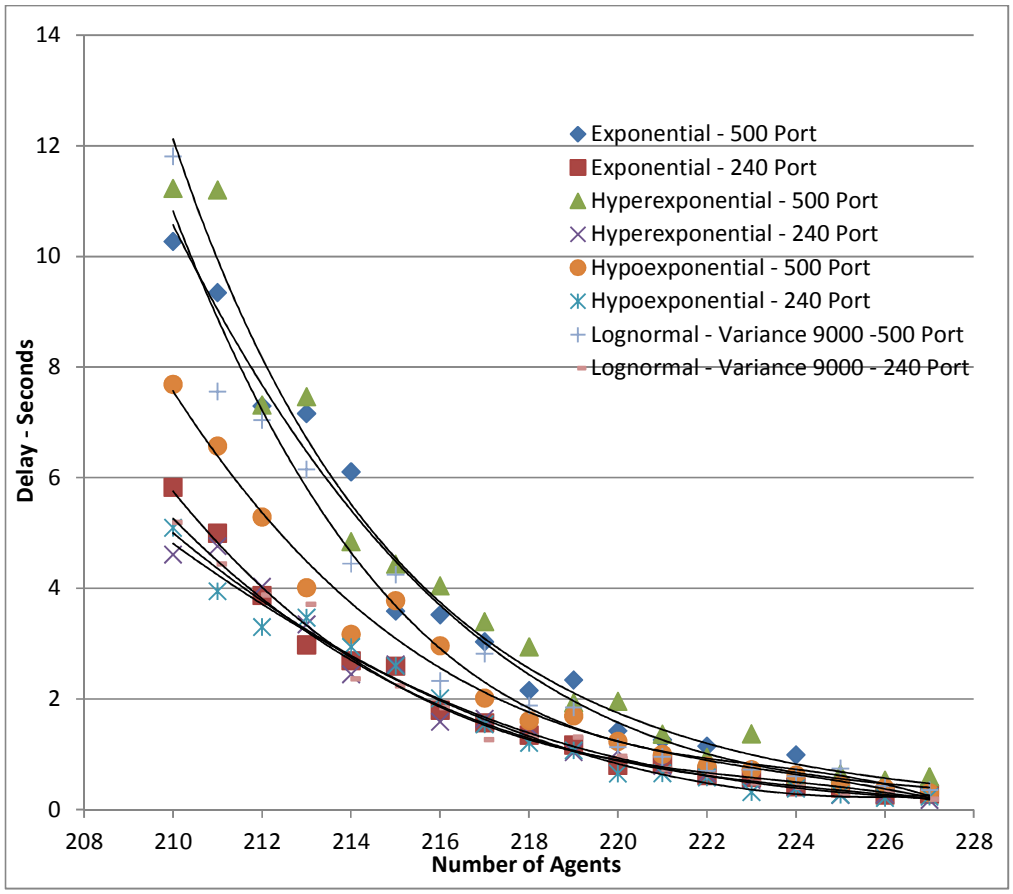

Figure 6 - Queuing delay comparison for different service time distributions

The results for the degree of queuing showed no distinction between the distribution selected; however, in the case of queuing times for the 500 port media gateway case we see some differentiation if service times are hypo exponentially distributed.

Many-server queues allow the utilisation of the individual servers to be high, whilst maintaining a relatively low degree of queuing within the overall system. Under some circumstances, in a heavy-load

regime [46], a steady state which is insensitive to the service-time distribution is seen. Light traffic theorems however appear to show that the insensitivity of blocking mechanisms to service time distribution is carried over into the analysis of queuing systems [47].

Hypo exponential distributions are the superimposition of two distributions in series, both of which have a mean that is less than that of the combined distribution. This is consistent with the observed lower queuing delays, seen for a hypo exponential distribution when there is no contention at the media gateway. This distinction is lost when we limit the queue size by imposing a very small degree of contention through a media gateway.

7. Cost Drivers

The description below illustrates how the results obtained from simulation may be used to form the basis of a cost benefit analysis. This has been built using the following assumptions based on experience and market rates:

- Service targets of: abandonment < 2% and 95% calls answered within 20s. A call cost of £0.05 per minute and an agent cost £17 per hour.
- A busy period, covering 8 hours of the business day.
- A peak volume of 2400 5 minute calls per hour is expected and a working week of 220 peak working days per year.
- A contract cost for the purpose of service credit calculation was calculated on the basis that agents are 70% of the cost of running the call centre.
- Service penalties, arising from failing to meet service targets, are 2.5% of the monthly fee. We assume that the worst case exposure is 3 months.
- Each lost or abandoned call means the loss of £50 worth of business. This is small when compared to the loss of large deals such as in stock trading, where every communication has a financial consequence.

This is not exact and gives a set of rough order of magnitude figures that allow the planner to consider the relative impact of service variables. The results are shown below:

Call Duration	5	Calls per hour	2400	Call Cost	0.05	per Minute	Busy Hrs	8	1.7	Overhead
				Agent	£17	per hour	Working Days	220	0.7	Agent Cost
Category	Agents to reach Service Level	Abandonment	Time to Abandon	Blocking	Call Cost	Cost Abandoned Calls	Agent Cost	Total Hourly Cost	Total Annual Cost	Annual total of Lost Calls
No Contention	214	0		0	597.3	0	£3,638	4235.3	7454128	0
	211	0.86%	300	0	594.5247	6.192	£3,587	4187.716735	7370381.453	36326.4
	210	1.86%	30	0	588.84	2.232	£3,570	4161.072	7323486.72	78566.4
Contention	213	0		0.45%	597.3	0	£3,621	4218.3	7424208	19008
	211	0.86%	300	0.05%	594.5247	6.192	£3,587	4187.716735	7370381.453	38565.12
	210	1.86%	30	0	588.84	2.232	£3,570	4161.072	7323486.72	78566.4
								Cost Variation	130641.28	
Revenue per lost call		50						Annual lost Revenue in lost calls	3928320	
Service Penalty Level		2.50%	3	Months		Estimated Contract Value	£15,549,851	Service Penalties	£97,186.57	

Table 2 – Outline call centre cost model

The figures (highlighted in grey) indicate that the variation in contract centre running costs, between the best and worst strategy, is about £130K per annum. However, the worst case figure for lost calls can lead to £4m in lost revenue; this varies depending on the value placed on a transaction. The potential exposure from service penalties is £97K per annum.

The impact on customer revenue is the biggest number and the hardest to quantify, minimising the number of lost calls is likely to be more important than cost minimisation.

The lowest operational cost is achieved when the highest abandonment level that will still meet the service performance criteria is observed. This however increases the number of lost calls and hence lost revenue. The difference in the number of lost calls between the 0.45% contention case and the highest permissible abandonment case is the latter loses around 595,584 calls more, which monetises to £2.98m per annum.

There have been several studies of the impact of the service parameters on performance. Feinberg [48] concludes that only two predictors of satisfaction, the percentage of issues resolved on the first call and average abandonment, show significant but low causality. Whilst the operators of call centres believed that First Call resolution, average speed of answer, blocked calls and finally abandonment are key drivers of customer satisfaction [49]; customers appear more concerned with the quality of the agent interaction [50].

8. Conclusion

Modelling the performance characteristics of a call centre in terms of the agent queue alone misses key performance influencers, specifically the interaction between channel availability at the media gateway and the time a call is queued. A blocking probability at the media gateway, as low as 0.45%, has a significant impact on the degree of queuing observed and therefore the cost and performance of the call centre. This fits with predictions from the extended Erlang C equation (2) and a revised evaluation of the waiting time using equation 8.

We have modelled the impact in the heavy traffic region and found that this follows the predictions of Erlang C and Whitt's heavy traffic approximation, with the impact of media gateway contention becoming visible when call volumes drop below agent capacity. We have also shown that the nature of the service time distribution has minimal impact on the behaviour seen, except in the case of the hypo exponential case due to the phases of the distribution being in series.

Abandonment also impacts queuing delays, however as the degree of abandonment reduces, the effect of contention at the media gateway continues to restrict the occupancy of the queue and limits the time a call can expect to wait. While the degree of abandonment has a very significant impact, the call centre manager has less control over this than he has over the level of contention at the media gateway. Our commercial assessment shows that in addressing the balance between abandonment and contention, we need to consider both operational costs and the loss of revenue due to lost calls.

There is the potential for extending this work to consider the impact of multi-skill call centres and ultimately the impact of multiple channels on queuing behaviour. We anticipate that there may be additional workload in passing wrongly directed calls to the right operator, while having more specialised operators reduces the overall number of operators per pool of specialism. In addition the impact of blocked customers or those who have elected to abandon a call, redialling is likely to be of interest, as traditionally this has been seen as weakening the validity of the Poisson call arrival assumption.

9. References

1. Gans, N., Koole, G., Mandelbaum, A.: 'Telephone call centers: a tutorial and literature review', Manufacturing and Service Operations Management, 2003, 5, (2), pp 79–141.
2. Sharp, D.: 'Call Center Operation: Design, Operation, and Maintenance', (Digital Press, Amsterdam, 2003).
3. Halfin, S. and Whitt, W.: Heavy traffic limits for queues with many exponential servers, Oper. Res. 1981 (29), pp. 567-588.
4. Whitt, W.: 'Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments', Management Science, October 2004, Vol. 50, Issue 10, pp 1449-1461.
5. Garnett, O., Mandelbaum, A., Reiman, M.: 'Designing a Call Center with Impatient Customers', Manufacturing & Service Operations Management. Summer 2002, Vol. 4, Issue 3, pp 208-227.
6. Whitt, W.: 'Sensitivity of performance in the Erlang-A queueing model to changes in the model parameters', Operations Research, 2006, Vol. 54, No. 2, pp 247–260.
7. Mandelbaum, A. and Zeltyn, S.: 'The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the M/M/n + G queue', OR Spectrum, July 2004, Vol. 26, Issue 3, pp 377-411.
8. Miller, L.E.: Formulas For Blocking Probability, https://www.yumpu.com/en/document/view/19394685/formulas-for-blocking-probability1-advanced-network-technologies- , accessed January 2017.
9. Jouini, O. and Roubos, A.: On multiple priority multi-server queues with impatience, Journal of the Operational Research Society; May 2014, Vol. 65 Issue 5, pp 616-632.
10. Jouini, O., Aksin, O.Z. and Dallery, Y.: Call centers with delay information: Models and insights, Manufacturing & Service Operations Management, Fall 2011, Vol. 13 Issue 4, pp 534-548.
11. Koole, G. and Pot, A.: A note on profit maximization and monotonicity for inbound call centers, Operations Research 2011, 59(5), pp 1304-1308.
12. Jouini, O. Koole, G.M. and Roubos, A.: Performance indicators for call centers with impatience. IIE Transactions 01/2012; DOI: 10.1080/0740817X.2012.712241.
13. Jouini, O.: Analysis of a last come first served queueing system with customer abandonment. Computers And Operations Research; 2012, 39(12): pp 3040-3045.
14. Mandelbaum, A. and Momcilovic, P.: Queues with Many Servers and Impatient Customers, OR, 2012, 37 (1), pp 41-65.

15. Ibrahim, R. and L'Ecuyer, P.: Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models,' Manufacturing and Services Operations Management, 2013, 15(1), pp 72-85.
16. Mandelbaum, A. and Zeltyn, S.: "The Palm/Erlang-A Queue, with Applications to Call Centers", June 19, 2005, http://ie.technion.ac.il/serveng/References/Erlang_A.pdf. Accessed September 2015.
17. Koole, G.M., Nielsen, B.F. and. Nielsen, T.B.: Optimization of overflow policies in call centers, Probability in the Engineering and Informational Sciences 2015, 29(3), pp 461-471.
18. Jaoua, A., L'Ecuyer, P. and Delorme, L.: Call-Type Dependence in Multiskill Call Centers, Simulation: Transactions of the Society for Modeling and Simulation International, 2013, 89, 6, pp 722-734.
19. Chan, W., Koole, G. and L'Ecuyer, P.: Dynamic Call Center Routing Policies Using Call Waiting and Agent Idle Times, Manufacturing and Service Operations Management, 2014, 16, 4, pp 544-560.
20. Chan, W., Ta, T. A., L'Ecuyer, P. and Bastin, F.: Chance-Constrained Staffing with Recourse for Multi-Skill Call Centers with Arrival-Rate Uncertainty, Proceedings of the 2014 Winter Simulation Conference, IEEE Press, 2014, pp 4103-4104.
21. Thiongane, M., Chan, W. and L'Ecuyer, P.: Waiting Time Predictors for Multiskill Call Centers, 2015 Winter Simulation Conference, Dec 2015, Huntington Beach, United States.
22. Whitt, W. and Perry, O.: Achieving Rapid Recovery in an Overload Control for a Large-Scale Service System. INFORMS Journal on Computing, Summer 2015, vol. 27, No. 3, pp. 491-506.
23. Koole, G., Legros, B. and Jouini, O.: Adaptive threshold policies for multi-channel call centers, IIE Transactions, 2015, 47(4), pp 414-430.
24. Koole, G., Roubos, A. and Stolletz, R.: Service level variability of inbound call centers, Manufacturing & Service Operations Management, 2012, 14(3), pp 402-413.
25. Whitt, W. and Kim, S.: Estimating Waiting Times with the Time-Varying Little's Law. Probability in the Engineering and Informational Sciences, 2013, vol. 27, pp 471-506.
26. Whitt, W. and Kim, S.: Statistical Analysis with Little's Law. Operations Research, July-August 2013, vol. 61, No. 4, pp 1030-1045.
27. Whitt, W.: Heavy-Traffic Limits for Queues with Periodic Arrival Rates *Operations Research Letters*, 2014, vol. 42, pp 458-461.
28. Whitt, W. and Perry, O.: Diffusion Approximation for an Overloaded X Model Via a Stochastic Averaging Principle. *Queueing Systems*, 2014, vol. 76, pp 347-401.

29. Whitt, W. and Perry, O.: A Fluid Limit for an Overloaded X Model Via a Stochastic Averaging Principle. Mathematics of Operations Research, May 2013, vol. 38, No. 2, pp 294-349.
30. Whitt, W. and Pang, G.: Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues with Dependent Service Times. Queueing Systems, 2013, vol. 73, No. 2, pp 119-146.
31. Whitt, W. and Liu, Y.: Many-Server Heavy-Traffic Limits for Queues with Time-Varying Parameters. *Annals of Applied Probability*, 2014, vol. 24, No. 1, pp. 378-421.
32. Whitt, W. and Liu, Y.: Algorithms for Time-Varying Networks of Many-Server Fluid Queues. *Informs Journal on Computing*, 2014, vol. 26, No. 1, pp. 59-73.
33. Whitt, W. and Kim, S.: Choosing Arrival Process Models for Service Systems: Tests of a Nonhomogeneous Poisson Process, Naval Research Logistics, 2014, vol. 61, No. 1, pp. 66-90.
34. Whitt, W. and Liu, Y.: Stabilizing Performance in Networks of Queues with Time-Varying Arrival Rates. Probability in the Engineering and Informational Sciences, 2014, vol. 28, No. 4, pp. 419-449.
35. Ibrahim, R., Ye, H., L'Ecuyer, P. and Shen H.: Modeling and Forecasting Call Center Arrivals: A Literature Survey, International Journal of Forecasting and a case study, July–September 2016, Volume 32, Issue 3, pp 865–874.
36. Roubos, A., Bhulai, S. and Koole, G.M.: Flexible staffing for call centers with non-stationary arrival rates, Markov Decision Processes in Practice (Richard Boucherie and Nico van Dijk, Eds), Springer, 2016.
37. Whitt, W.: Offered Load Analysis for Staffing. *Manufacturing and Service Operations Management*, Spring 2013, vol. 15, No. 2, pp. 166-169.
38. Koole, G., Jouini, O. and Roubos, A.: Performance indicators for call centers with impatience, IIE Transactions, 2013, 45(3), pp 341-354.
39. Ibrahim, R., L'Ecuyer, P., Shen, H. and Thiongane, M.: Inter-Dependent, Heterogeneous, and Time-Varying Service-Time Distributions in Call Centers', European Journal of Operational Research, 16 April 2016, Volume 250, Issue 2, pp 480–492.
40. Brown,L., Gans,N., Mandelbaum,A., Sakov,A., Shen,H., Zeltyn,S. and Zhao, L.: Statistical analysis of a telephone call center: A queueing science perspective, Journal of American Statistical Association, 2005,100(469), 36–50.
41. Feinberg, M.A.: Performance characteristics of automated call distribution systems. In GLOBECOM ’90, IEEE, 1990, pp 415–419.
42. Massey A. W. and Wallace B. R.: An optimal design of the M/M/C/K queue for call centers, See www.cs.cmu.edu/~harchol/WORMS04/talks/massey.ppt

43. Weerasinghe, A. and Mandelbaum, A.: Abandonment vs. Blocking in Many-Server Queues: Asymptotic Optimality in the QED Regime, QUESTA, 2013, 75 (2), pp 279-337.
44. Pitts, J.M and Schormans, J.A.: Configuring IP QoS Mechanisms for Graceful Degradation of Real-Time Services, MILCOM 2006 - 2006 IEEE Military Communications conference, 23-25 Oct. 2006, pp 1-7.
45. Li, A., Whitt, W. and Zhao, J.: Staffing To Stabilize Blocking In Loss Models With Time-Varying Arrival Rates, Probability in the Engineering and Informational Sciences , Volume 30 , Issue 02 , April 2016, pp 185 – 211.
46. Tschaikowski, M. and Tribastone,M.: Insensitivity to Service-time Distributions for Fluid Queueing Models, Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools, 2013, pp 273-281.
47. Burman, D.Y. and Smith, D.R.: Mathematics A light-traffic theorem for multiserver queues. Journal of Operations Research, Vol 8, No1, Feb 1983.
48. Feinberg, R., Kim, S., Hokama, L., de Ruyter, K. and Keen, C.: Operational Determinants of Caller Satisfaction in the Call Center, International Journal of Service Industry Management, 2000, Vol. 11(2), pp.131-141.
49. Boardman Liu, L.: Operationalizing Service Quality: Providers' Perspective, Northeast Decision Sciences Institute Proceedings, 2010, pp 533-8.
50. Boardman Liu, L.: Operationalizing Service Quality: Customers' Perspective, Proceedings for the Northeast Region Decision Sciences Institute; 2011, pp 1326-1331.
51. Mitrani, I., Probabilistic Modelling, Cambridge University Press, 1998.

10. Appendix

This section looks at the derivation of equation 8 in the main text.

If the waiting time distribution is dependent on not having to wait or having to wait for a time less than or equal to x, we can write for $x \geq 0$ [51]:

$$H(x) = P(Y_0 \leq x) = 1 - P_c + P_c P(Y_0 \leq x | Y_0 > 0)$$

If Y_0 is the target waiting time then the distribution function is given by the probability that the call does not have to queue plus the probability that it when it does, it queues for less than x. $P(Y_0 \leq x | Y_0 > 0)$ is

calculated as the sum of a geometrically distributed number (1-ρ) of independent and exponentially distributed random variables with parameter μ and distributed exponentially with parameter $(1-\rho)\,\mu$. In this case there are N servers (agents) and

$$(1-\rho)N\mu = \left(1-\frac{\lambda}{N\mu}\right)N\mu = (N\mu-\lambda)$$

So that

$$G(x) = P(Y_0 \le x | Y_0 > 0) = 1 - e^{-(N\mu-\lambda)x}$$

And

$$H(x) = 1 - P_c + P_c(1 - e^{-(N\mu-\lambda)x})$$

$$H(x) = 1 - P_c e^{-(N\mu-\lambda)x}$$

The average speed of answer (ASA), w is given by

$$w = E(x) = \int_0^\infty x dH(x) = \int_0^\infty x\, P_c\, (N\mu-\lambda)\, e^{-(N\mu-\lambda)x} dx$$

$$w = \left[-xP_c e^{-(N\mu-\lambda)x}\right]_0^\infty + \int_0^\infty P_c\, e^{-(N\mu-\lambda)x} dx$$

Yielding

$$w = \frac{P_c}{(N\mu-\lambda)}$$

Which after substituting for $\mu=1/T_s$ and $A=\lambda/\mu$ and $\rho=A/N$ becomes

$$w = \frac{P_c\ Ts}{(N-A)} = \frac{P_c\ Ts}{N(1-\rho)}$$

This is equation 3 in the main text, however suppose that we limit the number of ports into the system to R and therefore the maximum size of the queue becomes R-N and the maximum queueing time is given by

T_{max}=(R-N)/Nμ. Thus instead of integrating between the limits 0 and ∞ ; the integration should be done between the limits of 0 and $T_{max.}$ This gives the answer:

$$w = P_c\left(\frac{1}{(N\mu-\lambda)} - \left(\frac{1}{(N\mu-\lambda)} + T_{max}\right)e^{-(N\mu-\lambda)T_{max}}\right)$$

This can be re-expressed as:

$$w = P_c\left(\frac{Ts}{(N-A)} - \left(\frac{Ts}{(N-A)} + T_{max}\right)e^{-(N-A)T_{max}/T_s}\right)$$

This is the same as equation 8 in the main text.